

Definition-Only RAG for Financial QA: Average Gains and a Diagnostic Boundary

Leyi Ouyang Edison Marrese-Taylor Yutaka Matsuo

The University of Tokyo

leyi.ouyang@weblab.t.u-tokyo.ac.jp

Abstract

Previous work shows that a model’s ability to define domain terms correlates with performance in other tasks related to that domain. For example, in a QA setting, a language model can miss a question in a specialized domain for no reason other than not properly recalling what a term in the question means. We argue that a short glossary definition is the smallest unit of evidence a retrieval system can supply against that failure. Concretely, focusing on the finance NLP domain, we assess the role of definitions in improving performance on XFinBench Terminology-Understanding, and also pair MMLU Finance with a 13,609-term Investopedia glossary of financial terms. We experiment with a five-model panel spanning mid-capability open models to a frontier closed model, and show that definition-only retrieval produces a structured gain on these datasets, with the effect concentrating in a mid-capability band rather than scaling with model size. Finally, we connect term-level measurement to individual questions through a human-validated annotation, and find that term knowledge does not locate the questions where a definition changes the answer, with none of our implemented internal routing signals approaching the per-question oracle. We release the term-mapping annotation, the evaluation artifacts, and the analysis code that reproduces the load-bearing tables.¹

1 Introduction

Answering a question in a specialized domain often requires knowing what a particular term means, and a model that has not internalized that term can fail for that reason alone. For example, Qwen-2.5-7B answers a finance multiple-choice question with “American Depositary Receipt” where the correct answer is “Guilder Share”, a now-obsolete instrument with a similar function and a differ-

ent mechanism. Adding a one-sentence Investopedia definition of “Guilder Share” to the prompt flips the model to the correct option. Applied at scale, this kind of definition-only retrieval raises accuracy on multiple models in our panel across both MMLU Finance and XFinBench Terminology-Understanding.

Two observations motivate the study. Many finance benchmark questions can be answered only by knowing what a specific term means, and a domain glossary such as Investopedia stores exactly those meanings, so pairing the two suggests a narrow intervention in which we retrieve the term’s definition and inject it. The intervention comes with an equally narrow boundary, since it can help only where missing term knowledge, rather than reasoning or calibration, is what blocks the model. That boundary makes a second and more operational question natural, namely whether we can tell per question when retrieval will help and retrieve only there.

Definition modeling has been used as a probe of what a model has internalized about a domain since Noraset et al. (2017) and Gardner et al. (2022), with a financial extension by Jhirad et al. (2023). Marrese-Taylor et al. (2025) report that definition-modeling performance strongly correlates with downstream multilingual benchmark scores, which suggests the probe tracks something downstream tasks also depend on. Probe and intervention are therefore two uses of one resource, since the definitions a definition-modeling probe scores are exactly the glossary snippets a retriever injects. The signal that measures a model’s term knowledge could in principle also decide when retrieval is needed.

We accordingly ask two questions. We first assess whether retrieving short Investopedia definitions and injecting them into the prompt improves financial QA, and we then assess whether the same term-knowledge signal can tell us *when* retrieval should fire. Recent selective and adaptive RAG

¹<https://github.com/youngandlei/definition-only-rag>

work (Asai et al., 2024; Jeong et al., 2024) is motivated by the possibility that some observable property of the question or model state contains enough signal to route between retrieve and no-retrieve.

We find that definition-only retrieval helps on average while the term-knowledge signal does not predict when to retrieve. Across five models on both benchmarks the gains are uneven, with mid-capability open models improving and larger and frontier models not. We measure the same term-knowledge signal under both a surface metric and an LLM judge and connect it to outcomes through a per-question term annotation, and it sits at $|r| < 0.06$ against question correctness and against RAG gain for every model in the panel. Selective routing built on that signal, or on output-level confidence, recovers only a fraction of an oracle’s improvement, with only an external strong judge meaningfully exceeding always-on RAG. We describe this contrast as a *diagnosis–intervention gap*, in which term knowledge diagnoses where retrieval helps but does not tell us, per question, when to apply it.

2 Related Work

Definition modeling has been used as a probe of what a language model has internalized about a domain since Noraset et al. (2017) and Gardner et al. (2022), with a financial extension by Jhirad et al. (2023) and a multilingual extension by Marrese-Taylor et al. (2025). Recent selective and adaptive retrieval-augmented generation work (Lewis et al., 2020; Asai et al., 2024; Jeong et al., 2024) is motivated by the possibility that some observable property of the question or model state contains enough signal to route between retrieve and no-retrieve. We test the narrowest version of that assumption in finance, using definition-level term knowledge, which is the most semantically transparent candidate signal available in this domain. For uncertainty-style routing baselines we use semantic entropy (Farquhar et al., 2024) and option-margin signals, and we note that Wang et al. (2024) report that first-token MCQ scoring can disagree with the model’s final text answer, which limits how far option entropy alone can be trusted as a control signal.

3 Setup

Retrieval pipeline. A curated Investopedia term bank of 13,609 financial terms serves as both the

retrieval index and the reference set for term-level definition modeling. Each term-snippet representation concatenates the term name and its first three key-takeaway sentences. We use hybrid retrieval, combining a lexical retriever with a dense sentence-encoder, fusing the two ranked lists and reranking them with a cross-encoder, and we inject the top-3 reranked snippets above the question stem. The pipeline is fixed across all experiments.

Models and benchmarks. We evaluate five language models, namely Qwen-2.5-7B and 14B (Qwen Team, 2024), Llama-3.1-8B (Meta AI, 2024), Mistral-Small-24B (Mistral AI, 2025), and GPT-5.2, which we query in non-reasoning mode through the OpenAI Responses API. Open models are served through vLLM with logprob-based option scoring at temperature 0. Our main evaluation surface is MMLU Finance, the union of the finance-related MMLU subjects and 1,124 questions in total (Hendrycks et al., 2020), and we use XFinBench Terminology-Understanding (Zhang et al., 2025) as a 1,790-question secondary check on the same five models.

Term knowledge. For each (model, term) pair we prompt the model zero-shot for a definition, and we score each generation both with surface METEOR (Banerjee and Lavie, 2005) and with a 0–3 sense-correctness score from a GPT-4o-mini judge.

Per-question term annotation. The per-question analysis needs to know which glossary terms each question actually depends on, and no mapping between MMLU Finance and Investopedia exists, so we build one. Two LLM annotators (GPT-5.2 and GPT-4o) first map each MMLU Finance question to the glossary terms its correct answer depends on, through a two-stage pipeline that retrieves the dense top-30 candidates and then filters them by relevance with an LLM, and inter-annotator agreement between the two LLM annotators is $\kappa=0.76$, with mean Jaccard 0.73 and per-question F1 0.80 taking GPT-5.2 as reference. Because every per-question result downstream is read through this mapping, we do not treat LLM agreement as sufficient on its own, and we add a domain-trained human pass to make the connector trustworthy. The first author (undergraduate training in finance) reviewed all 2,233 (question, term) pairs from the union of the two LLM annotations against a written rubric

Model	Base	+RAG	Δ
<i>MMLU Finance</i> ($n=1,124$)			
Qwen-2.5-7B	60.8	64.1	+3.3
Qwen-2.5-14B	68.2	68.1	-0.1
Llama-3.1-8B	52.1	57.6	+5.4
Mistral-S-24B	72.3	72.5	+0.2
GPT-5.2	87.8	88.1	+0.3
<i>XFinBench TU</i> ($n=1,790$)			
Qwen-2.5-7B	70.5	75.3	+4.8
Qwen-2.5-14B	77.7	76.8	-0.8
Llama-3.1-8B	61.5	63.5	+2.0
Mistral-S-24B	78.9	78.9	± 0.0
GPT-5.2	87.6	88.5	+0.9

Table 1: Main panel: accuracy with and without RAG. **Bold Δ** : 95% CI excludes zero (paired bootstrap, 2000 resamples).

(Appendix B), keeping 84.7%, removing 15.3% as misalignments, and adding 20 further pairs (0.9%) where the reviewer identified load-bearing terms that both LLM annotators had missed. All per-question analyses run on the resulting human-validated annotation, which retains 995 term-bearing questions (averaging 1.90 validated terms each) and 129 (11.5%) zero-term controls. This mapping is the connector that ties term-level knowledge to per-question routing outcomes, and is released alongside the paper.

4 Definition-Only Retrieval Helps on Average

Table 1 reports baseline and RAG accuracies for the full five-model panel on both benchmarks. The largest gains land on the mid-capability open band, with Qwen-2.5-7B gaining +3.3 points and Llama-3.1-8B +5.4 on MMLU Finance, both significant under a paired bootstrap, while the other three models are not statistically significant. XFinBench TU reproduces the same shape on the same five models, with Qwen-2.5-7B (+4.8) and Llama-3.1-8B (+2.0) again the two significant rows and Qwen-2.5-14B (-0.8), Mistral-Small-24B (± 0.0), and GPT-5.2 (+0.9) all n.s.

If stronger models gained less because they already have the relevant terms parameterised, then the gain should concentrate in the subset where each model’s term knowledge is weakest, and we find that it does not. Sorting each model’s MMLU Finance questions into quartiles by the minimum TK-judge across annotated terms, Qwen-2.5-14B and GPT-5.2 are both at ± 0.0 points on their own Q1, and Mistral-Small-24B is +1.6, comparable to its aggregate. Stronger models are therefore not

failing on items where their own term knowledge is weakest, and adding a correct definition does not change their answers. Their residual errors come from reasoning, calibration, or decision-level mistakes, which retrieval does not address. Definition-only retrieval lifts accuracy where missing term knowledge is the constraint, and has no measurable effect where the constraint lies elsewhere.

A retrieved snippet could in principle raise accuracy by containing the answer string rather than by conveying the term’s meaning. Across all 1,124 MMLU Finance prompts (3,228 injected snippets), only 20 snippets contain a full answer-option string, and re-scoring with those snippets removed leaves accuracy identical (Table 2). Masking every term name inside its own snippet costs part of the gain but stays above the no-RAG baseline, so the snippet is doing work that the term name alone does not explain. Token overlap between snippet and options also fails to separate helpful from harmful retrievals, with newly chosen against abandoned option overlap at 0.033 against 0.046 on fixes and 0.044 against 0.045 on breaks.

Investopedia is almost certainly inside frontier pretraining data, which raises the converse worry that stalling models simply already hold the glossary, or that they ignore injected definitions altogether. We therefore test this with terms that cannot have been memorised. We write 50 synthetic financial terms with author-written definitions, each embedding one arbitrary discrete fact, and turn them into probe-style four-option MCQs with balanced answer positions and no term-name or option leakage. Without the definition both models sit near chance, and with it both answer every item correctly (Table 2). With memorisation impossible by construction, the frontier model exploits an injected definition as fully as the mid-capability one, so its flat Δ on MMLU Finance cannot be attributed to an inability or unwillingness to use retrieved definitions. This is also the deployment case in which new terminology postdates the training cutoff.

5 The Diagnosis–Intervention Gap

The gains in Table 1 come from always-on retrieval, which pays for retrieval on every question even though the snippet changes the answer on only a small subset. On Qwen-2.5-7B / MMLU Finance, only 222 of 1,124 questions are baseline/RAG disagreements at all, and they split into 102 questions retrieval fixes and 67 it breaks, with the remain-

(a) Leakage controls (Qwen-2.5-7B / MMLU)		
Condition	Acc.	vs RAG
No RAG (baseline)	60.8	-2.9
RAG (as reported)	63.7	—
drop answer-matching snippets	63.7	± 0.0
mask all term names	61.9	-1.8
(b) Invented terms ($n=50$, unmemorisable)		
Model	No def.	With def.
GPT-5.2	36.0	100.0
Qwen-2.5-7B	30.0	100.0

Table 2: Controls on the source of the gain. Removing snippets that contain an answer option changes nothing, and masking term names retains most of the gain. On terms no model can have memorised, both a frontier and a mid-capability model use the injected definition completely.

der changing between two wrong options. The aggregate +3.3 is therefore a residual of two-sided movement, not a uniform lift, and a per-question policy that kept the fixes while skipping the breaks would reach the 70.0% oracle of Table 5. A signal that flagged those 222 in advance, even imperfectly, would let a system retrieve selectively. Term knowledge is the most direct candidate, since it measures what the model has internalised about the very objects the pipeline delivers.

METEOR and the GPT-4o-mini judge rank our five models differently (GPT-5.2 is fifth under METEOR and first under the judge), because surface-overlap metrics systematically under-rate paraphrastic definitions. The metric choice can therefore affect any routing conclusion, and we accordingly test the per-question correlation under both views.

For each MMLU Finance question and each model we record (i) whether baseline is correct, (ii) whether RAG improves over baseline, and (iii) the model’s term-knowledge score on the question’s annotated terms. Because a question can depend on more than one term we aggregate in two ways, taking MEAN-K as the arithmetic mean across the question’s annotated terms and MIN-K as the worst-known term in the set, and the two correspond to different routing policies that trigger on overall unfamiliarity or on the weakest-known term respectively. The heatmap in Figure 1 summarises the resulting Pearson correlations under both aggregations and both scoring views, where r_{corr} is against baseline correctness and r_{Δ} against per-question RAG gain, over $n=995$ term-bearing questions each. The full numeric grid is in Ap-

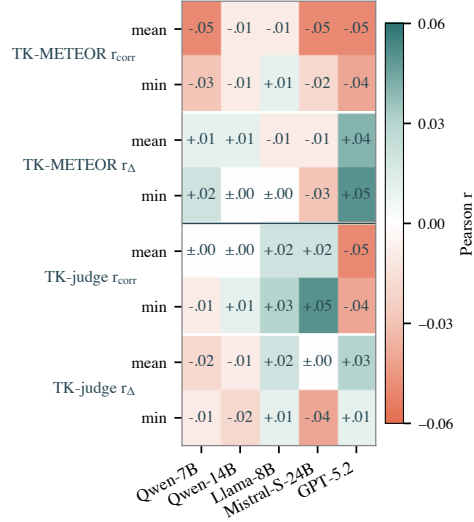


Figure 1: Correlation grid: term knowledge against outcomes. All forty correlations lie within ± 0.06 .

pendix F.

All forty correlations fall within $|r| < 0.06$ (max $|r|=0.053$), and each one’s 95% Fisher- z confidence interval covers zero. This is not a weak-evidence result but a tight two-sided bound, and it excludes a monotone signal strong enough to route on. A zero-term control (129 questions with no annotated term) shows no systematic accuracy difference from term-bearing questions, ruling out an annotation artifact. The absence of correlation is also not a retrieval failure in disguise: against the human-validated annotation the pipeline reaches recall@3 of 53% and hit-rate@3 of 74%, and restricting the grid to questions where an annotated term is actually among the injected top-3 leaves every correlation inside the same zero band, so the measured term is usually the term the model was given. A 1,000-trial permutation test shuffling term-knowledge labels across questions gives a reference distribution of max $|r|$ over the forty correlations with median 0.072 and 95% range $[0.047, 0.107]$; the observed max $|r|=0.053$ falls inside this range ($p=0.91$). Restricting to the subset of questions whose annotated terms fall in the same (or adjacent) per-model METEOR-quartile and judge-quartile (the subset where the two proxies measure consistently) leaves the picture unchanged (max $|r|=0.081$, Fisher- z bound ± 0.099 at $n=390$ –676 per model). At this scale, term knowledge is too weak to support a useful routing policy.

We test eight signals on Qwen-2.5-7B / MMLU Finance, of which seven are internal, namely response entropy, semantic entropy (Farquhar et al.,

Method	MMLU	XFin
Hard filter ($t=0.35$)	-0.80	-2.24
Adaptive top- k	-0.18	-0.78
Soft rerank ($\alpha=1$)	± 0.00	+0.16

Table 3: Definition-level selective injection (Δ points vs always-on RAG).

2024), calibrated confidence, retrieval-score thresholds, a learned trace-feature gate, a hidden-state probe, and same-model self-judging, and one is external, the GPT-4o-mini judge. Only the external judge meaningfully exceeds always-on RAG (67.4% vs always-on 64.1%, oracle 70.0%). Every internal signal underperforms always-on or recovers at most a third of the oracle gap, and threshold tuning on internal signals would not close it. Per-method accuracies in Appendix C.

Instead of gating retrieval on or off, we can keep it always on and filter *which* of the retrieved definitions to inject. We try three variants on Qwen-2.5-7B (Table 3), where hard METEOR filtering removes a snippet when the term’s METEOR is above a threshold and the model therefore already knows the term, adaptive top- k shrinks k when the retrieved terms are mostly well-known, and soft reranking keeps all retrievals but pushes lower-METEOR and hence less-known terms to the top. Hard filtering and adaptive- k lose ground on both benchmarks, and soft reranking is the closest to neutral.

If retrieval were strictly a knowledge-gap-filler, then removing definitions of already-known terms should be neutral, and we find it is consistently harmful instead, which indicates that definition snippets contribute something beyond filling missing knowledge and matches what we find per question.

To show what happens when the retrieved definition fully determines the answer, we generate 194 quantitative MCQs whose correct answer is a specific value embedded in the source term’s Investopedia definition, such as a formula result, a regulatory threshold, or a year, so that the retrieved snippet must carry the answer. Every MCQ was human-reviewed against a written rubric, and 192 of 194 generated items survive (review details in Appendix B). Δ_{RAG} lands at +21 to +33 points on the four open-source models (Mistral-Small-24B, at +0.2 on MMLU Finance, jumps from 61% to 94% here), an order of magnitude above the same models’ MMLU numbers, which is the diagnosis–

intervention gap in numerical form.

The two headline results rest on different evidence bases and we keep them apart accordingly. The average-gain pattern and the absence of per-question correlation both span the full five-model panel, the former replicated on two independently constructed benchmarks and the latter holding across all forty correlations in the model, measurement, and outcome grid. The result further survives restriction to the stricter both-annotator intersection (Appendix E). The routing sweep and the selective-injection experiments run on Qwen-2.5-7B / MMLU Finance alone, which is the panel’s deepest-instrumented setting and combines a significant gain, and hence measurable headroom, with the open weights that the hidden-state probe and self-judge signals require. We report them as evidence that cheap internal signals fail where the conditions for finding one are most favourable, not as a panel-wide result. Throughout, the claims are scoped to multiple-choice financial QA scored by option log-probability against a single curated glossary.

6 Conclusion

We have shown that definition-only retrieval improves MMLU Finance accuracy on mid-capability open models and reproduces the same pattern on XFinBench TU, with its reach bounded by the type of failure the model is making (§4). Our central finding is a *diagnosis–intervention gap*, in which the same term-knowledge signal that diagnoses where retrieval helps cannot decide, per question, when to apply it. No internal signal we test can decide it either, with an oracle showing a clear gain left on the table and only an external model capturing part of it. Closing this gap therefore needs information beyond the question and its definitions, plausibly the model’s own pretrained ability, where GPT-5.2’s near-saturation suggests the capability already resides. The controlled probe bounds the opposite regime, in which retrieval is reliably decisive. We release the term-mapping annotation together with the evaluation artifacts and analysis code.

Limitations

The knowledge source is a single domain-curated glossary, and the five-model panel covers one mid-capability open band, one mid-large open band, and one frontier closed model. A different reference

corpus or denser sampling of the 30–70B open range may change parts of the cross-model picture.

The per-question correlation analysis and the selective-routing experiments live on MMLU Finance, with Qwen-2.5-7B as the deepest-analysed model. XFinBench Terminology-Understanding ships its annotation against the XFinBench knowledge bank rather than against Investopedia, so cross-benchmark replication of the correlation analysis is left to future work.

The controlled set is built to be decisive, since its MCQs are written so that the correct answer is a specific value embedded in the retrievable definition. Its numbers are therefore a regime map for the upper bound of definition-only retrieval rather than a generic-benchmark result.

The evaluation is multiple-choice throughout. We chose that format because scoring is deterministic under option log-probabilities, which keeps paired per-question outcomes measurable at this sample size where free generation plus judge noise would blur them, and it also bounds what the results establish. Exam-style items test terminology more directly than real user requests do, and a fixed option set exposes the score competition that a definition perturbs, whereas open-ended answers can absorb a definition without any discrete outcome changing. Both the size of the average gain and the shape of the routing problem may differ for generative financial QA, since an open-ended setting has no equivalent of the four-state fix/break decomposition that bounds the oracle here, and a per-question routing signal would accordingly have to be defined against a graded rather than binary outcome. Extending the paired design to generative answers, and to questions drawn from real user traffic rather than curated exams, is the natural next step.

Acknowledgements

We thank the anonymous reviewers for their careful reading and constructive suggestions, one of which prompted the invented-terms control reported in §4. The experiments in this paper were run on GPU resources provided by the Matsuo Laboratory at the University of Tokyo.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. [Self-rag: Learning to retrieve, generate, and critique through self-reflection.](#)

In The Twelfth International Conference on Learning Representations.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization.*

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. [Detecting hallucinations in large language models using semantic entropy.](#) *Nature*, 630:625–630.

Noah Gardner, Hafiz Khan, and Chih-Cheng Hung. 2022. [Definition modeling: literature review and dataset analysis.](#) *Applied Computing and Intelligence*, 2(1):83–98.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. [Measuring massive multitask language understanding.](#) *arXiv preprint arXiv:2009.03300.*

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024. [Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity.](#) In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 7036–7050.

James Jhirad, Edison Marrese-Taylor, and Yutaka Matsuo. 2023. [Evaluating large language models’ understanding of financial terminology via definition modeling.](#) In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 93–100.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS).*

Edison Marrese-Taylor, Erica K. Shimomoto, Alfredo Solano, and Enrique Reid. 2025. [Multilingual definition modeling.](#) In *Findings of the Association for Computational Linguistics: ACL 2025.*

Meta AI. 2024. The Llama 3 herd of models. Technical report, Meta.

Mistral AI. 2025. Mistral small 3 (24b instruct). Model release, [mistral.ai/news/mistral-small-3.](#)

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085.*

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. [Definition modeling: Learning to define word embeddings in natural language](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Qwen Team. 2024. Qwen2.5 technical report. Technical report, Alibaba Cloud.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. [“My Answer is C”: First-token probabilities do not match text answers in instruction-tuned language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416.

Zhihan Zhang, Yixin Cao, and Lizi Liao. 2025. [Xfin-bench: Benchmarking llms in complex financial problem solving and reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2025*.

A Comparison with Jhirad et al. 2023 under matched scoring

Jhirad et al. (2023) define two scoring protocols for definition modelling on Investopedia: a strict protocol (first Key Takeaway only) and a lenient protocol (max METEOR over the first KTAW plus all subsets of the remaining KTAWs). Their main reported table uses the lenient protocol; we re-evaluate the 2026 panel under both to make the comparison protocol-matched.

Model	Params	Strict	Lenient
<i>Jhirad et al. 2023 (re-tabulated)</i>			
OPT-IML	1.3B	15.92	—
GPT-J	6.0B	22.17	—
GPT-JT	6.0B	6.65	—
MT0-XXL	13B	24.68	—
FLAN-UL2	20B	27.30	—
GPT-3.5-Turbo	175B	35.82	—
<i>Our 2026 extension</i>			
Qwen-2.5-7B	7B	30.82	34.71
Qwen-2.5-14B	14B	31.04	34.29
Llama-3.1-8B	8B	28.98	32.35
Mistral-Small-24B	24B	31.86	34.40
GPT-5.2	—	25.27	27.41

Table 4: Definition-modelling METEOR ($\times 100$) vs Jhirad et al. (2023) (strict / lenient).

Under either protocol, the 2026 panel does not exceed GPT-3.5-Turbo on this surface metric (strongest 2026: Mistral-Small-24B 31.86 strict / Qwen-2.5-7B 34.71 lenient). GPT-5.2 is lowest (25.27 / 27.41), consistent with the METEOR-versus-judge rank reversal in §5.

B MMLU Finance Term Annotation: Details and Release

Pipeline. For each MMLU Finance question we (i) generate 30 candidate Investopedia terms using a dense retriever over the question stem and answer choices, then (ii) prompt GPT-5.2 to keep only the candidate terms whose definitions are required to answer the question correctly. We run the same pipeline a second time with GPT-4o as an independent annotator over all 1,124 questions.

Inter-annotator agreement. Cohen’s κ between the two LLM annotators is 0.76 on the binary keep/discard label. Set-level mean Jaccard is 0.73, and per-question F1 with GPT-5.2 as the reference annotator is 0.80. The exact-match rate on the kept set is 54%.

Annotation density. The validated annotation retains 995 term-bearing questions averaging 1.90 terms per question, with 129 (11.5%) zero-term controls used in §5. Per-subject density is highest on quantitative subjects (HS Macroeconomics 1.86 mean terms/q over 390 questions; HS Microeconomics 1.59 / 238; Professional Accounting 1.45 / 282) and lowest on reasoning-driven subjects (Econometrics 1.12 / 114; Business Ethics 1.03 / 100).

Human review pass. After the LLM annotation, the first author (undergraduate training in finance) reviewed every (question, term) pair against a written rubric: a pair is retained iff (i) the term’s Investopedia definition directly supports the correct answer, (ii) removing the term would make the answer ambiguous to a domain-trained reader, and (iii) the term contributes to the answer’s semantic construction (not just its surface form). Pairs failing any criterion were removed; pairs missed by both LLM annotators but satisfying all three were added.

Edit-rate summary over 2,233 LLM-proposed (question, term) pairs: 84.7% kept, 15.3% removed as misalignments. The human reviewer further added 20 (question, term) pairs (0.9% of the LLM-proposed base) where a load-bearing term had been missed by both LLM annotators. The retention rate stratifies sharply by LLM agreement: of pairs that both GPT-5.2 and GPT-4o picked ($n=1,495$), 98.1% were retained; of pairs picked only by GPT-4o ($n=490$), 58.2% were retained; of pairs picked only by GPT-5.2 ($n=248$), 56.9% were retained. Agreement therefore calibrates the

human pass rather than the choice of union: it tells the reviewer which pairs need scrutiny, since single-annotator proposals are close to a coin flip while agreed pairs are almost always sound. The union is what maximises recall, and discarding the 738 single-annotator proposals would have cost the roughly 57% of them that survived review; the human pass is what restores precision. The small but non-zero set of human-added pairs shows that even the union leaves residual gaps (Appendix E).

Definition Probe review. The same author reviewed every Definition Probe MCQ against a written rubric requiring (i) the correct answer to be verbatim or directly derivable from the source definition, (ii) type-consistent choices, (iii) no distractor admitting an alternative correct answer, and (iv) no surface leak of the source term into a choice. Of 194 generated items, 191 were retained as-is; 1 had a distractor revision (a quoted-spread question whose canonical answer was missing from the choices); and 2 were dropped because the source-term name leaked into a choice (“25% Rule” and “Half-Life”).

Release format. The released annotation is a per-question CSV (*question_id*, *gold_inv_term_id*), shipped alongside a companion file with the term names resolved, the second-LLM-annotator CSV, the stricter both-annotator intersection, the written rubric, and the agreement report. The controlled probe ships its reviewed items, the source terms, the drafting prompt, and the construction notes.

C Selective Routing: Per-Method Results

Table 5 reports accuracy for every routing method discussed in §5 on Qwen-2.5-7B / MMLU Finance. All accuracies are over the full $n=1,124$ set; the policies differ only on the 222 baseline/RAG disagreement questions. The pattern is dominated by one positive entry (GPT-4o-mini judge at 67.4%, recovering more than half the oracle gap from 64.1% toward 70.0%) and a long tail of internal-signal methods that match or slightly exceed always-on RAG without meaningfully closing the oracle gap.

Two further methodological notes. (i) The learned logistic gate and the hidden-state probe are evaluated under stratified 5-fold cross-validation with no test labels used to fit either classifier. (ii) The retrieval-score and calibrated-confidence thresholds are swept on the full set and we report

Method	Accuracy
Baseline	60.8
Always-on RAG	64.1
Entropy gate	61.4
Semantic entropy [†]	61.4
Calibrated confidence margin	62.3
Retrieval-score threshold	63.4
Hidden-state probe ($\Delta L21$, PCA128)	65.4
Learned gate (LR, trace features)	64.2
Self-judge (Qwen-2.5-7B)	64.9
GPT-4o-mini judge	67.4
Oracle (baseline vs RAG, per question)	70.0

Table 5: Routing-signal accuracies (Qwen-2.5-7B / MMLU Finance). [†]Farquhar et al. (2024).

the best resulting accuracy; this upper-bounds a held-out cutoff, which biases the comparison *in favour* of these methods. The conclusion that they still do not exceed always-on RAG is correspondingly stronger.

D Reproducibility

Retrieval pipeline. Lexical side: scikit-learn TF-IDF over the term snippets, English stop words, 50k max features. Dense side: sentence-transformers/all-MiniLM-L6-v2 (Reimers and Gurevych, 2019), a 384-dim bi-encoder. The query is the question stem concatenated with the four choice strings. Each retriever returns its top-12; the two lists are merged by reciprocal-rank fusion ($k=60$), and the fused top-12 is reranked by the cross-encoder cross-encoder/ms-marco-MiniLM-L-6-v2 (Nogueira and Cho, 2019). The top-3 reranked snippets are injected. A snippet is the term name followed by its leading Key-Takeaway sentences. The pipeline is identical in every experiment; only the presence of the injected block varies.

Prompt template. The question and its four options come first, then a Definitions: block holding the top-3 snippets, then Answer:. We refer to this as the definitions-last template. Scoring takes the argmax over the four option-letter log-probabilities at temperature 0, so decoding is deterministic and no answer parsing is involved. Confidence intervals are paired bootstrap over questions with 2,000 resamples.

Benchmark composition. MMLU Finance is the union of five MMLU subjects: econometrics (114), high_school_macro-economics (390), high_school_micro-economics (238), profes-

sional_accounting (282), and business_ethics (100), for 1,124 questions. XFinBench TU is the Terminology-Understanding split (1,790 questions), built by a different group on a different question pool and used as a replication surface.

Routing-signal implementations. Response entropy and option margin are computed from the four baseline option log-probabilities (Shannon entropy; top-1 minus top-2). Semantic entropy samples $k=5$ generations and clusters them by propositional equivalence before computing entropy. Calibrated confidence applies temperature scaling fitted on a held-out split; scaling cuts calibration error from 0.087 to 0.060 without changing routing accuracy, so the margin’s failure is not a calibration artifact. The retrieval-score threshold gates on the reranker’s top-1 score. The learned gate is a logistic regression over 27 logit, entropy, and retrieval trace features. The hidden-state probe is a logistic regression over the layer-21 RAG-minus-baseline state difference reduced to 128 dimensions by PCA; because it reads that difference, it cannot decide whether to retrieve before retrieving. Both learned methods use stratified 5-fold cross-validation. The two threshold methods are swept on the full set and reported at their best value, which upper-bounds a held-out cutoff and biases the comparison in their favour.

Term identifiers. The Investopedia resource of Jhirad et al. (2023) is not publicly hosted; we obtained it from the authors and keep their nomenclature, where a term ID is the dataset’s top-level integer index (for example 759 Confidence Interval, 3748 Null Hypothesis, 4839 Sharpe Ratio). The release ships the full ID-to-term mapping alongside the annotation so that no external resolution step is needed.

Glossary scope. The annotation is glossary-relative by design, since the intervention can only retrieve what the glossary contains. Questions whose load-bearing knowledge falls outside Investopedia (for example “fixed effects model”) carry no annotated term and enter the analysis as zero-term controls.

E Validation Studies

Blind human audit of the LLM judge. To check that the sense-correctness judge is not itself producing the per-question result, the first author blind re-scored 200 generated definitions, stratified over

judge levels and models, on the judge’s own 0–3 rubric without seeing its scores. Exact agreement is 66.5%, within-one-point agreement 99.0%, and agreement on the binary sense-correct split 94.0%. Where the two disagree the human is more often the more generous of the pair (48 cases vs 19), so the judge errs conservative; a uniformly conservative judge lowers term-knowledge scores across the board and cannot manufacture a spurious absence of correlation.

Independent from-scratch re-annotation. The main annotation has a human reviewing LLM-proposed candidates, which bounds recall by what the LLMs proposed. To estimate what that misses, an economics PhD who is not an author selected terms from scratch for 50 randomly sampled MMLU Finance questions, blind to the LLM candidates and to the released annotation. Of the rater’s 34 in-glossary pairs, 22 were already present, 8 are near-duplicates or fail the necessity criterion, and 4 are load-bearing pairs the pipeline missed (for example “M1” on a question asking what M1 includes). This is a 5.4% pair-level miss rate against the 0.9% estimated within the review protocol, so independent selection does surface a small tail. Missed pairs act as attenuating noise on the correlation, which moves the estimate toward zero rather than away from it, and they perturb under 1% of the annotation. The recovered pairs are folded into the release.

Intersection-annotation robustness. The main pipeline takes the union of the two LLM annotators and restores precision by human review, which maximises recall: an intersection would discard all 738 single-annotator proposals, of which about 57% survived review. As a robustness check we also ran the stricter variant. On both-annotator-only pairs ($n=959$ questions) the full grid of forty correlations is unchanged, with $\max |r|=0.05$ and every one inside the zero band. The choice of union over intersection therefore does not drive the result.

F Per-Question Correlation Grid

Table 6 gives the full numeric grid behind the heatmap in §5: per-question Pearson correlations on MMLU Finance, under both MEAN-K / MIN-K aggregations and both scoring views. The Fisher- z 95% CI half-width at $r=0$ is ± 0.062 , so every correlation’s individual CI includes zero.

Model	TK-METEOR		TK-JUDGE	
	r_c	r_Δ	r_c	r_Δ
	mn/mi	mn/mi	mn/mi	mn/mi
Qwen-2.5-7B	-.04/-.02	+.01/+.02	+.04/+.03	-.03/+.00
Qwen-2.5-14B	-.02/+.00	+.00/-.02	-.02/+.01	+.00/-.02
Llama-3.1-8B	-.04/-.03	+.03/+.02	-.00/+.01	+.05/+.03
Mistral-S-24B	-.06/-.05	-.02/-.03	-.00/-.00	+.00/-.01
GPT-5.2	-.05/-.03	+.03/+.04	-.00/-.01	+.02/+.00

Table 6: Per-question correlation grid: term knowledge vs outcomes.